



Biocomputing and HPC

Dr. Manfred Zorn

March 20, 2001
Old Dominion University
Norfolk, VA



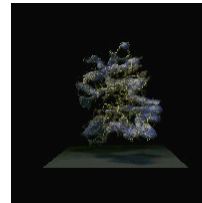
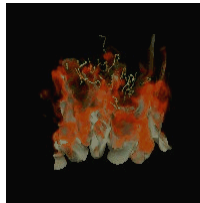
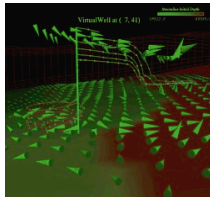
Lawrence Berkeley National Laboratory



- Founded in 1931 by Ernest O. Lawrence
- Best known for Particle Physics, found a dozen new transuranic elements: Bk, Cf, Am, Lw, Pu, ..., Sg
- About 4000 people, 800 students, 2000 visitors
- National User Facilities:
 - Advanced Light Source
 - NERSC Supercomputing Center

Old Dominion University

- **the Department of Energy, Office of Science, supercomputing facility**
- **unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines**
- **25th anniversary in 1999**



Old Dominion University

- **NERSC-3 (IBM SP3/RS 6000)**
- **Phase I: June 1999**
 - 608 processors
 - 410 gigaflop peak performance
 - Provides one teraflop NERSC capability
- **Phase II: December 2000**
 - 2,432 processors
 - 3.2 teraflop peak performance
 - 4 teraflop total NERSC capability



Old Dominion University



Center for Bioinformatics and Computational Genomics



■ Vision

- **A national center for understanding information and information systems in modern biology**

■ History

- **Established July 1998 within NERSC at LBNL by merging the Bioinformatics Group and the Human Genome Field Office**
- **Co-directed by Sylvia Spengler and Manfred Zorn**



Old Dominion University



Center for Bioinformatics and Computational Genomics



■ Research

- **Special Analysis Tools: Fold Prediction, Phylogeny, genome comparisons**
- **Compute-intensive Algorithms: clustering, phylogeny**

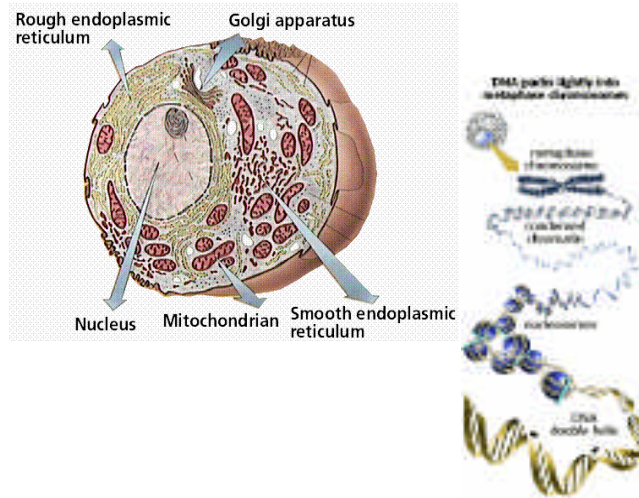
■ Development and Support

- **Large-scale Genome Annotation**
- **Wet lab support for Biologists**

■ Public Service

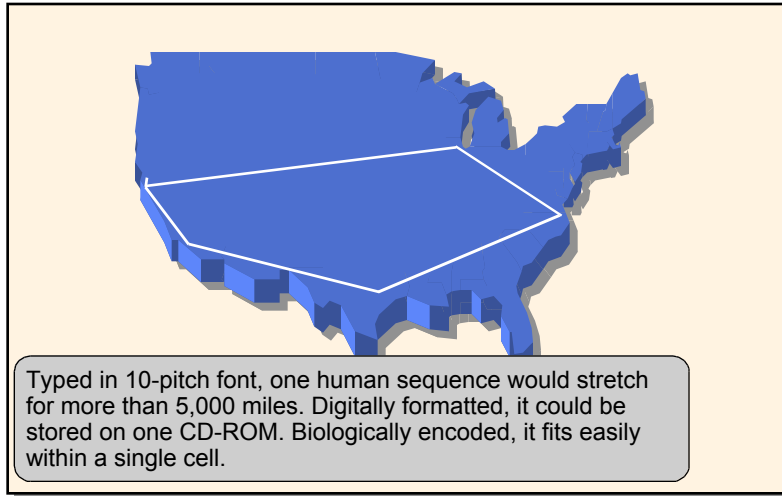
- **Public databases**
- **Education and Outreach, Standards**

Old Dominion University



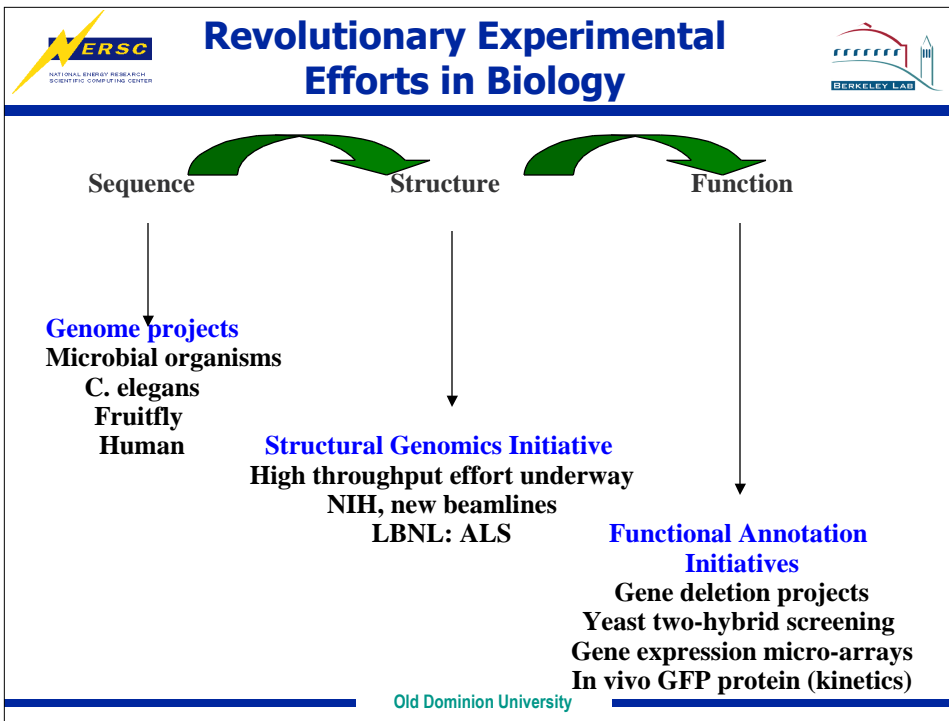
- **24 Chromosomes**
 - ✓ 1 – 22, X, Y
 - ✓ 23 pairs
- **1 Mitochondrial Genome**
- **3 Billion Base Pairs**
- **~30,000 Genes**

One Human Sequence



Genome Projects

1995 H. influenzae	2 Mb
1996 S. cerevisiae	12 Mb
1997 E. coli	5 Mb
1998 C. elegans	100 Mb
1999 Human Chromosome 22	34 Mb
2000 D. melanogaster	140 Mb
2000 H. sapiens	3,000 Mb



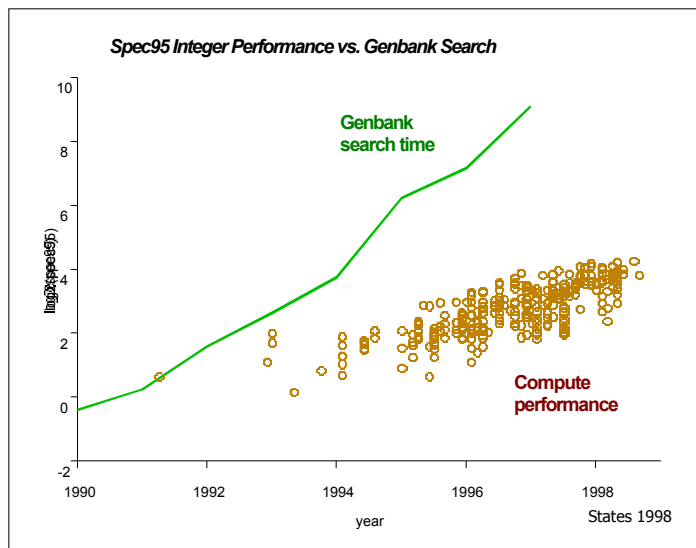
High-Throughput Genome Sequence Assembly, Modeling, and Annotation

The Genome Channel Browser to access and visualize current data flow, analysis and modeling. (Manfred Zorn, NERSC)

- Genome sequencing and annotation → Bioinformatics
- 30,000 human genes; genes from other organism
- Structure/functional annotation at the sequence level
- Computation to determine regions of a genome that might yield new folds
- Experimental Structural Genomics Initiative
- Functional annotation at the structure level by experiment

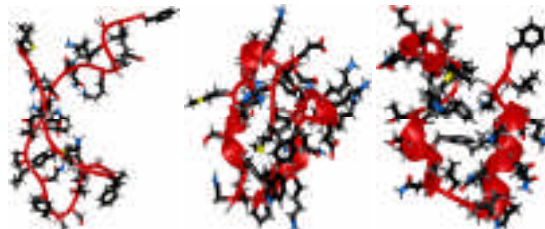
Old Dominion University

Moore's Law and Genomics



Old Dominion University

Low Resolution Fold Topologies to High Resolution Structure



One microsecond simulation of a fragment of the protein, Villin. Duan & Kollman, Science 1998

Low Resolution Structures from Predicted Fold Topology

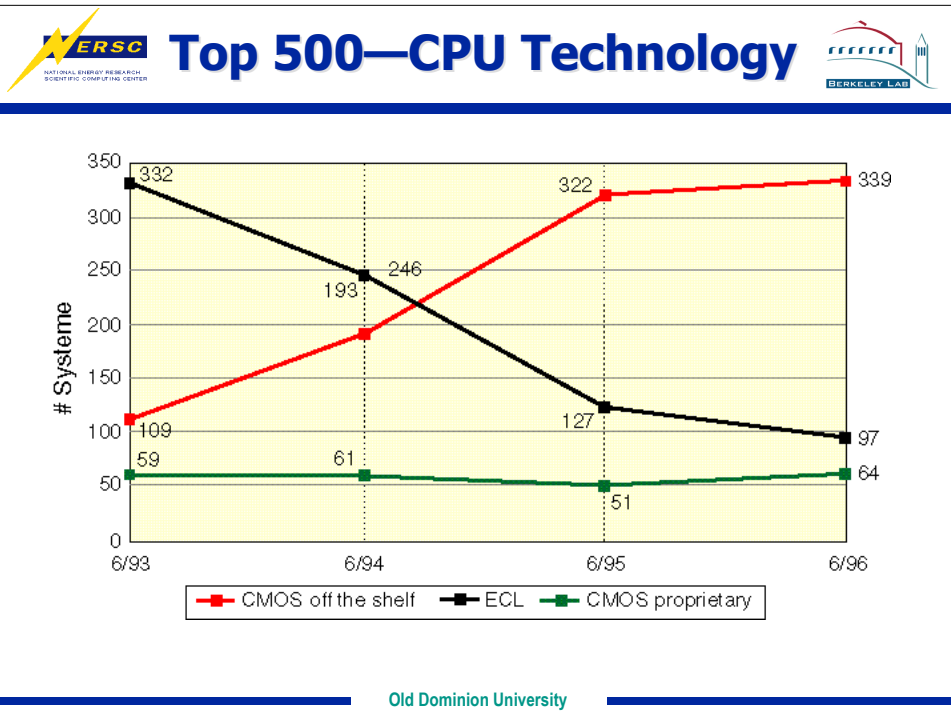
Fold class gives some idea of biological function, but....



Higher Resolution Structures with Biochemical Relevance

Drug design, bioremediation, diseases of new pathogen

Old Dominion University



ERSC What's supercomputing got to do with genomics? **BERKELEY LAB**

- Complexity of the information
- Amount of data
- Most applications are trivially parallel

Old Dominion University



The Need for Advanced Computing for Computational Biology



Computational Complexity arises from inherent factors:

At least 30,000 gene products just from human; many more from other organisms

Experimental data is accumulating rapidly

N^2 , N^3 , N^4 , etc. interactions between gene products

Combinatorial libraries of potential drugs/ligands

New materials that elaborate on native gene products from many organisms

Algorithmic Issues to make it tractable

Objective Functions

Optimization

Treatment of Long-ranged Interactions

Overcoming Size and Time scale bottlenecks

Statistics

Old Dominion University



DNA Sequencing



Read base code from storage medium!

■ **Read length: About 600 bases at once**

■ **Reader capacity**

✓ 100 lanes in parallel in about 5 hours

✓ 100 lanes in parallel in about 2 hours

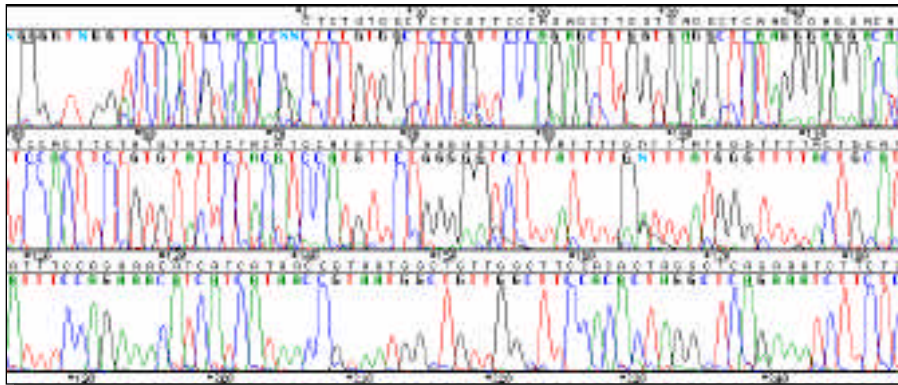
3 Billion year old program store

Old Dominion University

Sequencing: "bird's eye view"

- **Prepare DNA**
 - about a trillion DNA molecules
- **Do the sequencing reactions**
 - synthesize a new strand with terminators
- **Separate fragments**
 - by time, length = constant
- **Sequence determination**
 - automatic reading with laser detection systems

Sequence Traces



Good quality sequence needs
about 10X Coverage



Any genome is larger than amount of sequence that can be generated in a single step.

- **Shotgun**
- **Directed**
- **Finishing**

- **Break DNA into manageable pieces**
- **Sequence each piece**
- **Use sequence to reassemble original DNA**

Uniform process
Easily automatable





Assembly

Putting humpty-dumpty together again!

- **Overlap**
 - ✓ Find overlapping fragments
- **Layout**
 - ✓ Order and orientation of fragments
- **Consensus**
 - ✓ Determining the consensus sequence
- **Use of constraints**

Old Dominion University

- Repeats,
 - repeats,
 - ✓ repeats,
 - ◆ Repeats
 - ◆ 200 bp Alu repeat every ~4,000 bp with 5% -15% error
- Clipping
- Orientation
- Contamination
- Rearrangements
- Sequencing errors
- True Polymorphisms

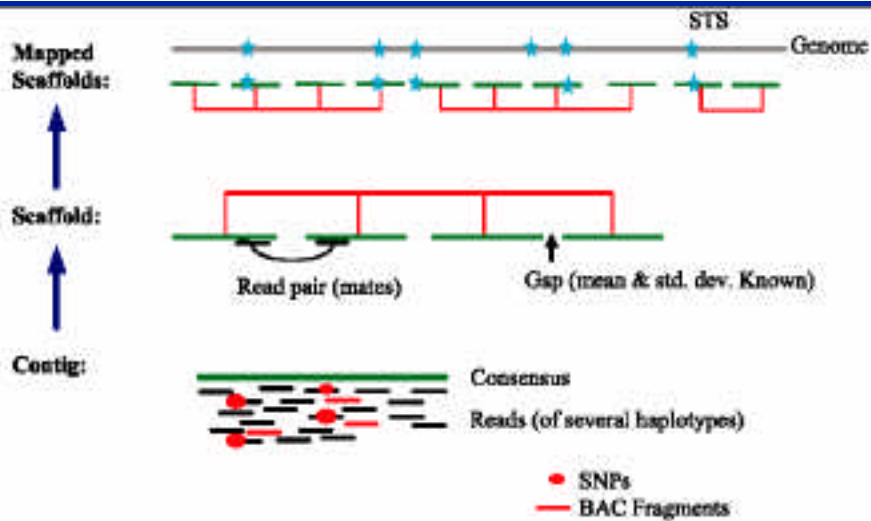
- Break DNA into manageable pieces
- Map pieces into tiling path
- Repeat

Two separate processes: mapping and sequencing
 More difficult to automate
 Hard to integrate map information into assembly
- 1.



- Use maps to assemble original DNA

- Special cases that drop out of the pipeline
- Gap closing
- Difficult stretches
- Primer walking
- Different strains, vectors, chemistry
- Creative solutions,





Whole Genome Assembly



8:37	Screeners
86:25	Overlapper
38:29	Unitiger
4:12	Scaffolder
5:44	Repeat Resolution I,II
25:05	Consensus

■ Human Genome

- ✓ 20,000 CPU hours
- ✓ (10,000 for overlapper)

Old Dominion University



DNA Analysis



- Heuristics
- Statistics
- Artistics

Old Dominion University

Disassemble the base code!

- Find the genes
 - Heuristic signals
 - Inherent features
 - Intelligent methods
- Characterize each gene
 - Compare with other genes
 - Find functional components
 - Predict features

- **Definition:** An inheritable trait associated with a region of DNA that codes for a polypeptide chain or specifies an RNA molecule which in turn have an influence on some characteristic phenotype of the organism.

Abstract concept that describes a complex phenomenon

What is Annotation?

- **Definition:** Extraction, definition, and interpretation of features on the genome sequence derived by integrating computational tools and biological knowledge.

Identifiable features in the sequence

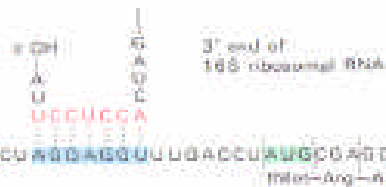
Heuristic Signals

**DNA contains various recognition sites
for internal machinery**

- **Promoter signals**
- **Transcription start signals**
- **Start Codon**
- **Exon, Intron boundaries**
- **Transcription termination signals**

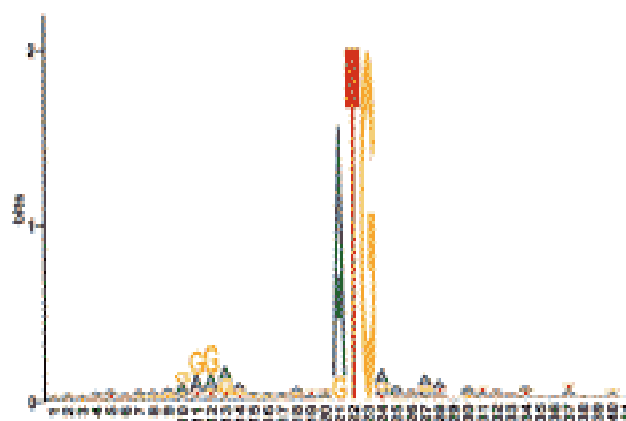
Initiation

AGCAC CAGGG AAAAUCUGAUG GAACGCUAC	<i>E. coli</i> <i>hsp4</i>
UUUGGAT GGAUUGAAACGALUGCGAUUGCA	<i>E. coli</i> <i>sdhB</i>
GGUATC CAGGUAACAACCAUCCGAGUUGUG	<i>E. coli</i> <i>rib4</i>
CAAUUCAGGUGGUGAUAUCUAAACCAAGUA	<i>E. coli</i> <i>rib7</i>
AAUCUUGAGGCUUUUUUAUGGUUCUUUUCU	ϕ X174 phage ϕ protein
UAAC UAAGGUGAUAUUGCAUCUCUAAAGACA	Gf phage replicase
UCCUAGGAGGUUUGACCUAUGCGAGCUUUU	R17 phage λ protein
AUGUAC UAAGCAGGUUGUAUGGAACAACGC	λ phage <i>am</i>



Old Dominion University

Start Codon



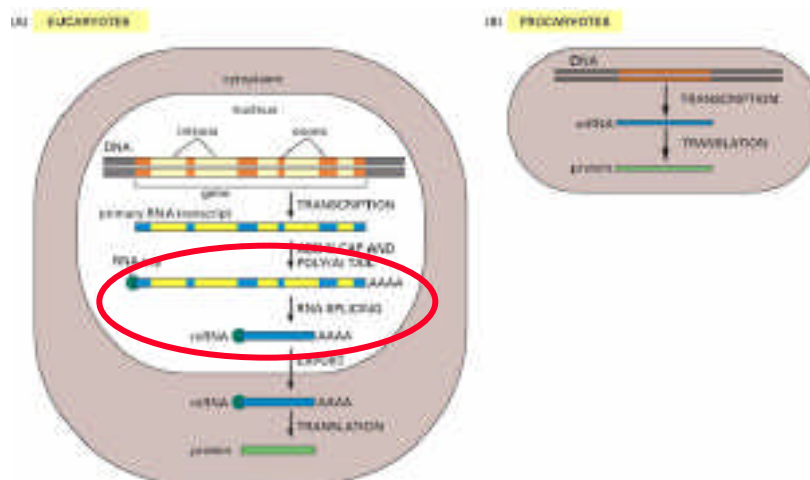
Old Dominion University

Start of the gene

Old Dominion University

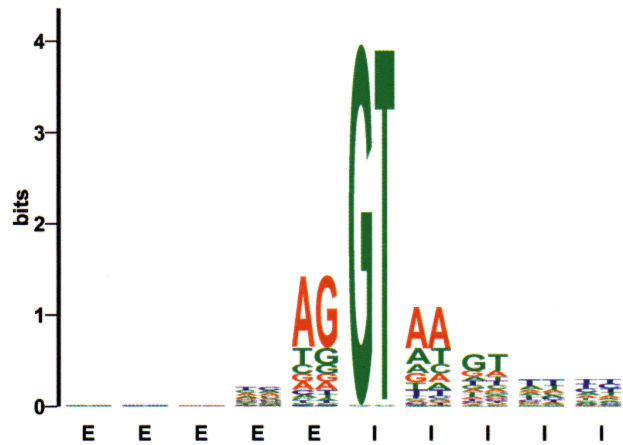
DNA exhibits certain biases that can be exploited to locate coding regions

- Uneven distribution of bases
- Codon bias
- CpG islands
- In-phase words
- Encoded amino acid sequence
- Imperfect periodicity
- Other global patterns



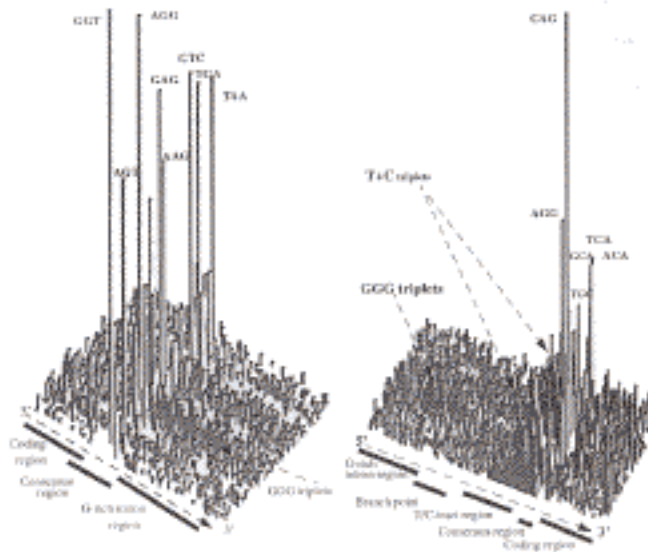
Donor Splice Site

R. Durbin et al. 1999



Old Dominion University

Inherent Features

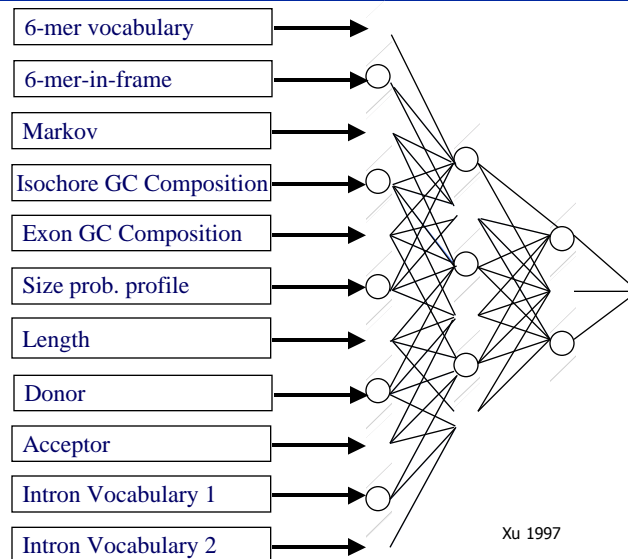


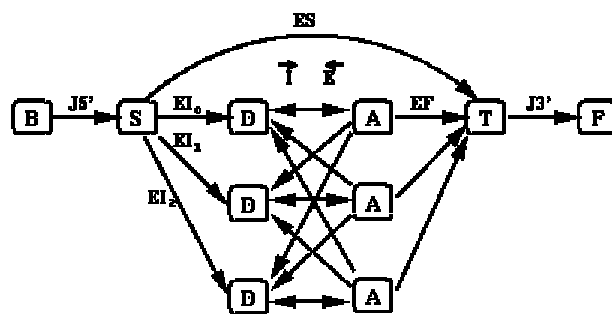
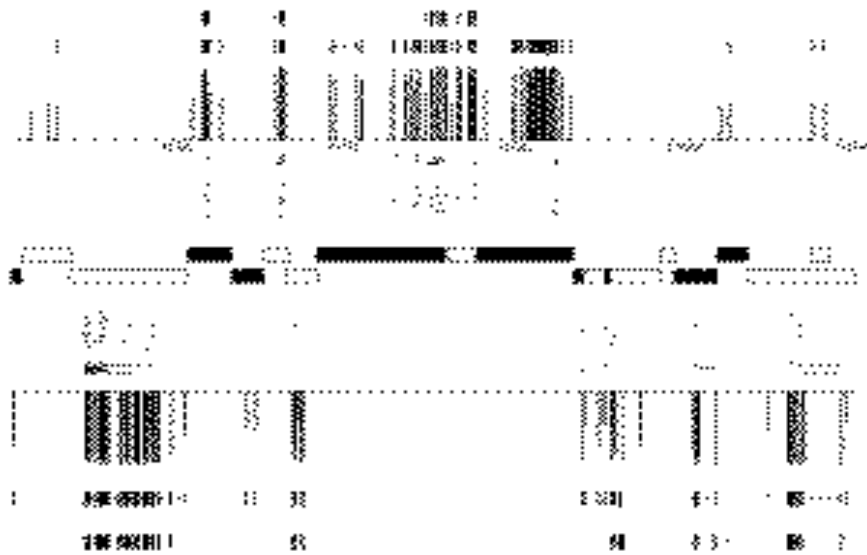
Solov'yev, 1994


Old Dominion University

Pattern recognition methods weigh inputs and predict gene location

- Neural Networks
- Hidden Markov Models
- Stochastic Context-Free Grammar






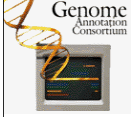


ERSC
NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER

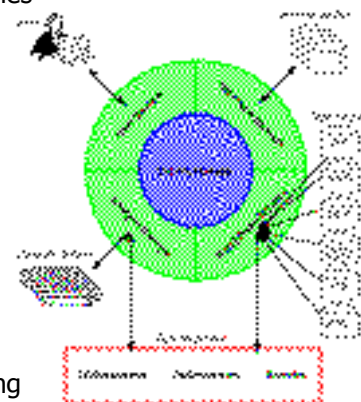
Large-scale Genome Annotation



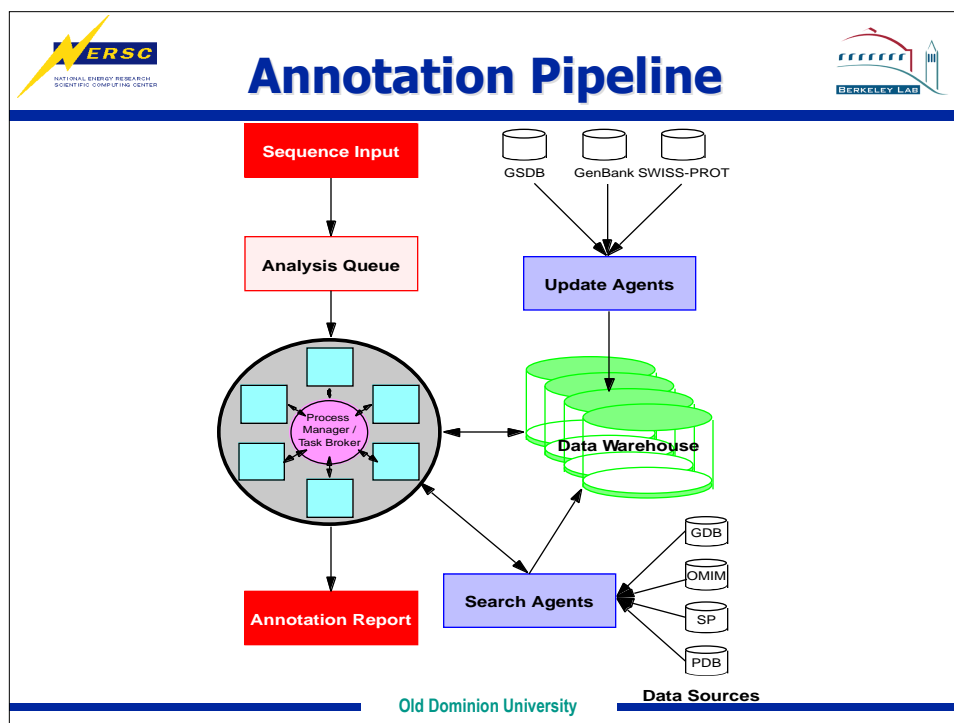
BERKELEY LAB

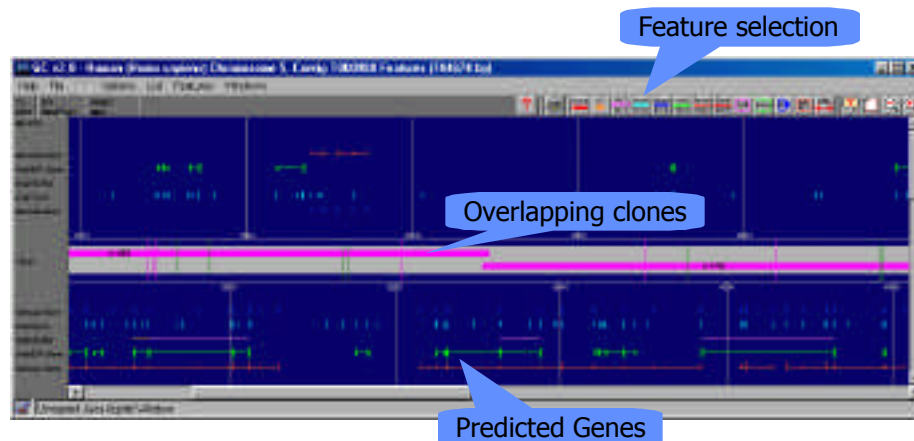
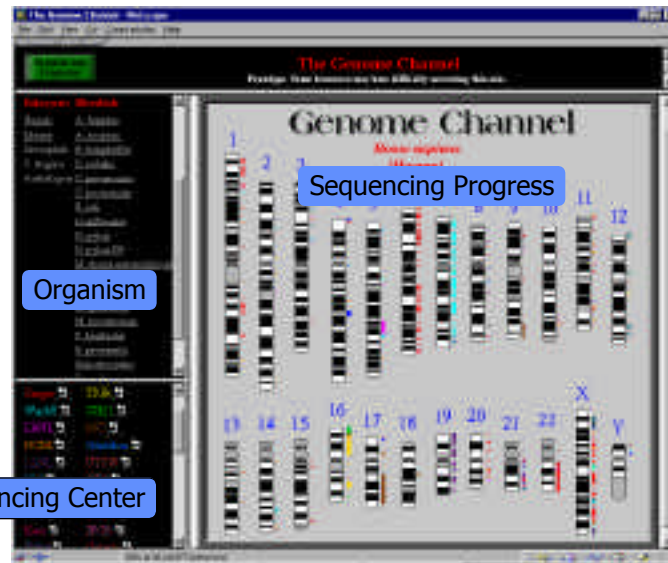
 **Genome Annotation Consortium**

- Multi-laboratory Project
- Standard Annotation of Genomes
 - Genome Channel
 - Genome Catalog
- Comprehensive integration of
 - Analysis tools
 - Data management systems
 - Data mining
 - User services
- Extensible Framework
 - High-performance computing
 - Data integration technology
 - Artificial intelligence



Old Dominion University







- **Distributed processes on many machines**
 - Workstation and Linux clusters
 - Opportunistic use of high performance computers
- **Approximately weekly update of:**
 - State of genome assemblies, clones and contigs
 - All genome sequence analysis
 - Done for all included genomes
 - Re-evaluate links to related data /gene and protein reports



Sequence Analysis Toolkit



- **Grail EXP .. (Uberbacher, Mural, Hyatt, Xu)**
- **Genscan .. (Burge and Karlin)**
 - ORNL workstation cluster implementation of code
- **Genie .. (David Haussler, in progress)**
 - HMM-based gene modeler
 - ORNL workstation cluster implementation of code
- **Microbial Genomes**
 - Microbial GRAILs
 - Glimmer, Genmark, Critica (Genbank)
- **tRNAscan .. (Sean Eddy)**

Old Dominion University



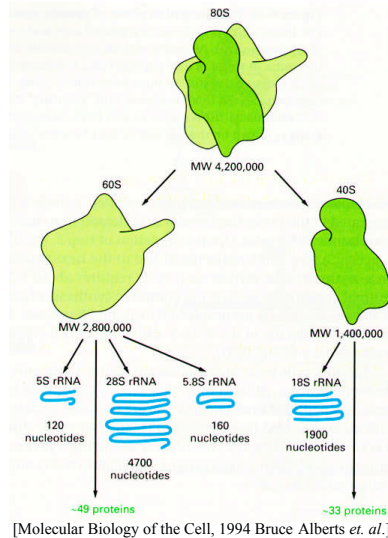
CPU Requirements



- **Last year's Genome annotation**
 - 250 Mbases DNA yield ~125 Gbytes of data
 - It takes ~ 7.5 days on 20 workstations ~3,600nhr
- **Celera's Fruitfly Sequencing**
 - Assembly of 1.7 Million reads in 25 hrs
 - Annotation 8-10 Mbases per months with 6 FTE
- **Celera's Human Sequencing**
 - 26.4 Million reads, 14.4 Billion base pairs, 4.6X
 - Assembly of Human Genome:
20,000 CPU hours on 160 Alpha-Processor Compaq cluster in about one month

Old Dominion University

Use of Ribosomal RNA (rRNA) in Phylogenetic Analysis



- Ribosomes are the subcellular machinery responsible for protein production
- Ribosomes ubiquitous, with similar structures, functions
- rRNA major constituent of ribosomes
- rRNA highly conserved among organisms
- rRNA good candidate molecule for evolutionary studies

Old Dominion University

Ribosomal RNA Project II

Secondary Structure, small subunit, ribosomal RNA

- **Ribosomal Database Project II**
Center for Microbial Ecology
Michigan State University
- **National Energy Research Scientific Computing Division, LBNL**
- **Indiana University, IN**

- **Semi-automatic alignment**
- **Value-added microbiology**
- **Real world limitations**
- **Characterization of microbial populations**
- **Isolation of new organisms**
- **Shorter and more sequences**

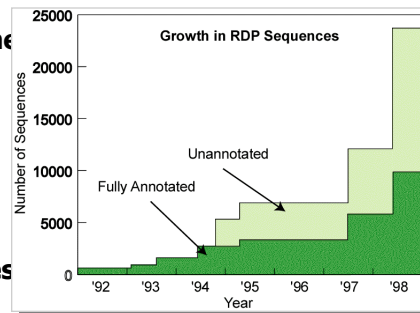
Old Dominion University



<http://www.cme.msu.edu/RDP>

Maidak et al, Nucleic Acids Res. 27:1171-1173 (1999)

- Value-added database with aligned rRNA sequences and annotation
- Analysis and visualization tools
- Reducing the lag time for harvesting sequence data from GenBank
- Improving annotation procedures



- Phylogenetic tree calculations on NERSC Supercomputer
 - Porting existing software
 - Improvements to increase throughput and capacity
 - New clustering algorithm



Reconstructing history from DNA sequences



- DNA changes over time; much of this change is not expressed
- Changes in unexpressed DNA can be modeled as Markov processes
- By comparing similar regions of DNA from different organisms (or different genes) one can infer the phylogenetic tree and evolutionary history that seems the best explanation of the current situation

Old Dominion University



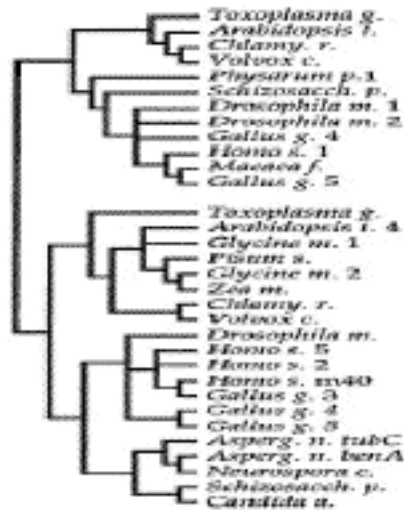
Changes in genetic information over time



- **Point mutations**
DNA – sequences of the 4 nucleotides
CCTCTGAC
vs
TCTCCGAC
Protein – sequences of the 20 amino acids
GSAQVKGHGKK
vs
GNPKVKAHGKK
- **Insertions and deletions**
DNA
CCTCT+GAC
vs
CCTCTTGAC

Old Dominion University

Why is tree-building a HPC problem?



- The number of bifurcating unrooted trees for n taxa is $(2n-5)! / (n-3)! 2n-3$
- for 50 taxa the number of possible trees is ~ 1074 ; most scientists are interested in much larger problems
- The number of rooted trees is $(2n-5)!$

Alignment

- To build trees one compares and relates 'similar' segments of genetic data. Getting 'similar' right is absolutely critical!
- Methods:
 - dynamic programming
 - Hidden Markov Models
 - Pattern matching
- Some alignment packages:
 - BLAST
<http://www.ncbi.nlm.nih.gov/BLAST/>
 - FASTA
<http://gcg.nhri.org.tw/fasta.html>
 - MUSCA
<http://www.research.ibm.com/bioinformatics/home>

Matching cost function

GCTAAATTC

++ x x

GC AAGTT

- Penalize for mismatches, for opening of gap, and for gap length
- This approach assumes independence of loci: good assumption for DNA, some problems with respect to amino acids, significant problems with RNA

Phylogenetic methodologies

- Define a specific series of steps to produce the 'best' tree
 - Pair-group cluster analyses
 - Fast, but tend not to address underlying evolutionary mechanisms
- Define criteria for comparing different trees and judging which is better. Two steps:
 - Define the objective function (evolutionary biology)
 - Generate and compare trees (computation)
- All of the techniques described produce an unrooted tree.
- The trees produced likewise describe relationships among extant taxa, not the progress of evolution over time.

Distance-based Tree-building methods

- **Aligned sequences are compared, and analysis is based on the differences between sequences, rather than the original sequence data.**
- **Less computationally intensive than character-based methods**
- **Tend to be problematic when sequences are highly divergent**

Distance-based Tree building methods, 2

- **Cluster analysis.** Most common variant is Unweighted Pair Group Method with Arithmetic Mean (UPGMA) – join two closest neighbors, average pair, keep going. Problematic when highly diverged sequences are involved
- **Additive tree methods** – built on assumption that the lengths of branches can be summed to create some measure of overall evolution.
 - **Fitch-Margoliash (FM)** – minimizes squared deviation between observed data and inferred tree.
 - **Minimum evolution (ME)** – finds shortest tree consistent with data
- **Of the distance methods, ME is the most widely implemented in computer programs**



Character-based methods



- Use character data (actual sequences) rather than distance data
- **Maximum parsimony.** Creates shortest tree – one with fewest changes. Inter-site rate heterogeneity creates difficulties for this approach.
- **Maximum likelihood.** Searches for the evolutionary model that has the highest likelihood value given the data. In simulation studies ML tends to outperform others, but is also computationally intensive.

Old Dominion University



fastDNAmI



- Developed by Gary Olsen
- Derived from Felsensteins's PHYLIP programs
- One of the more commonly used ML methods
- The first phylogenetic software implemented in a parallel program (at Argonne National Laboratory, using P4 libraries)
- Olsen, G.J., et al. 1994. fastDNAmI: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Computer Applications in Biosciences 10: 41-48
- MPI version produced in collaboration with Indiana University will be available soon

Old Dominion University



fastDNAmI algorithm



- Compute the optimal tree for three taxa (chosen randomly) - only one topology possible
- Randomly pick another taxon, and consider each of the $2i-5$ trees possible by adding this taxon into the first, three-taxa tree.
- Keep the best (maximum likelihood tree)
- Local branch rearrangement: move any subtree to a neighboring branch ($2i-6$ possibilities)
- Keep best resulting tree
- Repeat this step until local swapping no longer improves likelihood value

Old Dominion University



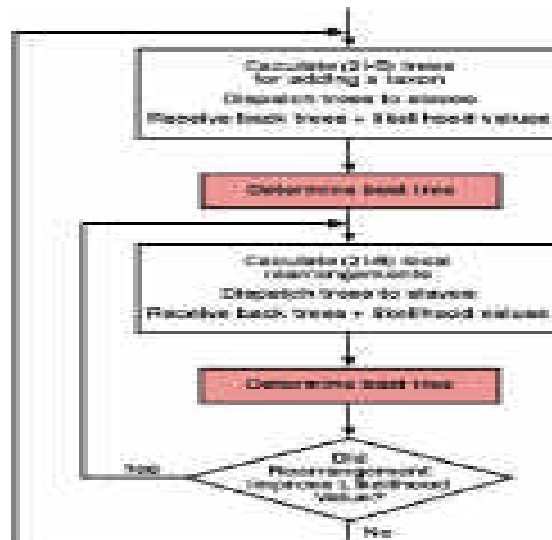
fastDNAmI algorithm con't: Iterate



- Get sequence data for next taxon
- Add new taxa ($2i-5$)
- Keep best
- Local rearrangements ($2i-6$)
- Keep best
- Keep going....
- When all taxa have been added, perform a full tree check

Old Dominion University

Overview of parallel program flow



Old Dominion University

Because of local effects....

- Where you end up sometimes depends on where you start
- This process searches a huge space of possible trees, and is thus dependent upon the randomly selected initial taxa
- Can get stuck in local optimum, rather than global
- Must do multiple runs with different randomizations of taxon entry order, and compare the results
- Similar trees and likelihood values provide some confidence, but still the space of all possible trees has not been searched extensively

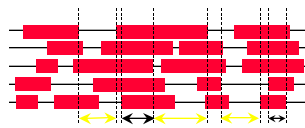
Old Dominion University

Stage I: Sequence segmentation

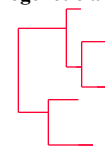
sequence collection



multiple alignment



clustering or
phylogenetic analysis



Segment: set of consecutive columns in alignment

GOAL: optimally partition multi-alignment into k maximally homogenous segments

TECHNIQUE: analog of *image processing* procedure
Statistical profile info + dynamic programming

Stage II: Clustering on segments

Divide segment into core cluster and 'tail'

Find $H^* = \arg(\max F(H))$

Minimum split function (measures compactness):

$$F_x(H) := \min_{i \in H} \pi(i, H).$$

Internal linkage function $\pi(i, H)$ (measures similarity btw i, H):

$$\pi(i, H) = |y_i| - \alpha |Y^H|, \quad Y^H = \bigcup_{i \in H} y_i$$

Generally: **optimal clustering procedure is exponential. For a monotonic linkage function, there is a polynomial optimization procedure (Muchnik et al, 1997).**

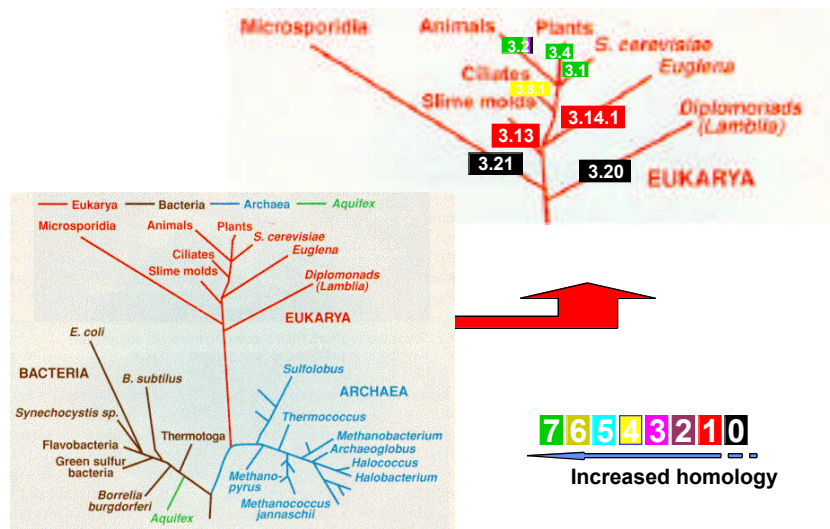
Cluster Intersection Results

Eukaryotes: Out of $2^7 = 128$ possible distribution patterns, the 409 sequences we analyzed, fall into only 33 patterns.

Among these 33 patterns, only **four** show significant frequencies, accounting for 81% of the total:

```
[0000000]: number_of_sequence=32
[0011111]: number_of_sequence=14
[1111110]: number_of_sequence=48
[1111111]: number_of_sequence=249
```

Compare cluster rank with phylogenetic tree*



* [E. Pennesi, *Science* 280, 672-674, 1998]

Possible Interpretations

- Different types of primitive organisms have more heterogeneous genetic background
- Higher organisms have more shared genetic material
- Higher organisms (i.e., the metazoan phylum) develop functional diversity of the genes based on this shared genetic background

Credits

- NERSC / LBNL
 - Donn Davy
 - Inna Dubchak
 - Sylvia Spengler
 - Eric P. Xing
 - Manfred Zorn
- ORNL
 - Ed Uberbacher
 - Richard Mural
 - Phil LoCascio
 - Sergey Petrov
 - Manesh Shah
 - Morey Parang
- Indiana University
 - Craig Stewart